www.reglament.net/bank/credit/2005/2/statya_1.htm. ***3.*** Беляев И.П. Принцип инвариантности лидерства. Электронный научный журнал «Исследовано в России» //http/zhurnal.ape.relarn.ru/articles/2006/115.pdf.

Поступила в редколлегию 15.11.07

УДК 004.8

Т. І. ЗАВАЛІЙ, аспірант кафедри ІСМ, НУ «Львівська політехніка»,
Ю. В. НІКОЛЬСЬКИЙ, канд. фіз.-мат. наук, доцент кафедри ІСМ,
НУ «Львівська політехніка»

ЯКІСНІ ХАРАКТЕРИСТИКИ МОДЕЛЕЙ ПРИЙНЯТТЯ РІШЕНЬ, ОТРИМАНИХ ІЗ ВИКОРИСТАННЯМ НАБЛИЖЕНИХ МНОЖИН

Предметом дослідження є результати психологічного тестування, які використовують для прийняття рішень щодо допуску до роботи операторів енергетичних мереж. Описано підхід, який використовує наближені множини для пошуку у таблицях даних правил, за якими приймають відповідні рішення. Виведені правила використано для класифікації нових прикладів. Розглянуто питання оцінювання якості класифікаторів за допомогою ROC-кривих і коефіцієнта успішності. Аналіз здійснено програмною системою Rosetta.

Предметом исследований являются результаты психологического тестирования, которые используют при принятии решения о допуске к работе операторов энергетических сетей. Для поиска правил принятия решений в таблицах данных предлагается использовать приближенные множества. С помощью полученных правил выполнена классификация новых примеров. Расматривается оценивание качества классификаторов с помощью ROC-кривых и коефициента успешности. Анализ произведен программной системой Rosetta.

The authors analyzed psychological testing results, which were used for a decision-making about admitting the power engineering specialists to work. The rough sets approach and its application for mining rules from data tables is described. These rules were used to classify new examples. The role of success rate and ROC-curve in a classifier's evaluation is considered. All of the analysis work was done with Rosetta software.

Вступ. Методи машинного навчання та інтелектуального аналізу даних широко використовують для пошуку закономірностей в даних, побудови моделей прийняття рішень, класифікації та кластеризації прикладів. Інтелектуальний аналіз − це процес побудови моделей прийняття рішень на основі даних, або виведення шаблонів з даних. Побудовані моделі відіграють роль корисного знання, викристалізованого з даних. Більшість методів інтелектуального аналізу базуються на методах машинного навчання, розпізнавання образів, математичної статистики. Еволюція та використання цих методів на даний час є областю активних досліджень [1, 2].

Одними з найпоширеніших в інтелектуальному аналізі є м'які обчислення [3] у складі нечітких множин, наближених множин, генетичних алгоритмів, нейронних мереж тощо. Вони оперують такими поняттями, як неточність, шум, надлишковість, непевність, наближене логічне виведення. Методи м'яких обчислень показують хороші результати у вирішенні реальних задач, обробці великих обсягів зашумлених даних, моделюванні складних предметних областей.

Проблема, результати вирішення якої наведено у цій статті, полягає в аналізі даних психологічних тестувань. Ці дані використано для побудови правил прийняття рішень. Відповідні тестування проводились на підприємствах енергетичної галузі з метою моніторингу та покращення роботи персоналу. Особливістю аналізованих даних є те, що емпіричні дані тестувань, зібрані психологами, доповнені суб'єктивними оцінками керівництва. В зв'язку з цим, виникає проблема об'єктивного оцінювання й оптимізації тестування персоналу, яку вирішено аналізом і застосуванням побудованих моделей прийняття рішень.

Наближені множини. Теорія *наближених множин* (rough sets) [4, 5, 6] вперше була сформульована Ж. Павлаком (Z. Pawlak). Вона є математичним інструментом подолання шуму та надлишковості в даних на основі використання понять верхнього та нижнього наближень, граничної області, розрізненності прикладів тощо. Цю теорію застосовують в поєднанні з іншими методами на всіх етапах видобування знань.

Алгоритми на базі теорії наближених множин використовують для зменшення кількості атрибутів у таблицях даних та побудови правил вигляду "якщо-то". Для цього спочатку формують таблицю прийняття рішень $\mathbf{A}=(U,A\cup\{d\})$, де U – непорожня скінченна множина прикладів, A – непорожня скінченна множина умовних атрибутів, d – атрибут прийняття рішення з доменом V_d, $|V_d| = k$. Значення v_i атрибута d відносить кожний приклад $x\in U$ до класу прийняття рішення X_i, де $i = 1,2,\ldots,k$. Із таблиці **A** застосуванням спеціальних методів аналізу видаляють зайві атрибути та знаходять *редукт* – мінімальну множину атрибутів, яка зберігає початкову розрізненність прикладів і залежності, наявні в таблиці. Одним з методів пошуку редукта є *логічне виведення* (boolean reasoning). Цей метод дозволяє шукати наближені розв'язки шляхом регулювання *ступеня підтримки* (HF, hitting fraction) редукта. Використання ступеня підтримки є одним із механізмів боротьби з шумом в даних та подолання проблеми *перенавчання* (overfitting), оскільки наближений редукт містить лише найважливіші атрибути та відображає найсильніші залежності в таблиці прийняття рішень. Однією з реалізацій логічного виведення є алгоритм Джонсона [7]. Після знаходження редукта генерують правила прийняття рішень та обчислюють їхні якісні характеристики. Правила утворюють модель прийняття рішень, яку використовують для класифікації нових прикладів. Цю модель назвемо класифікатором.

Класифікація. Класифікатор – це четвірка $C = \langle RUL, P, HF, \tau \rangle$, де *RUL* – множина правил, *P* – характеристики правил, τ – *порогове значення*, яке є дійсним числом, $\tau \in [0, 1]$. У загальному випадку класифікацію виконують за методом голосування [8]. За цим методом на основі параметрів *P* розраховують коефіцієнти впевненості в належності приклада до кожного з наявних класів. Клас із найбільшим значенням коефіцієнта впевненості "перемагає" в класифікації приклада. У разі бінарної класифікації, якщо коефіцієнт впевненості для одного з двох наявних класів перевищує задане порогове значення τ, то приклад відносять саме до цього класу.

Якість правила $\alpha \to \beta$ (*якщо* α, *то* β) оцінюють за такими числовими характеристиками [8]:

1. Підтримка, *support*($\alpha \to \beta$) – кількість навчальних прикладів, для яких виконуються як умова α правила, так і його наслідок β.
2. Точність, *accuracy*($\alpha \to \beta$) – відношення кількості навчальних прикладів, для яких виконується все правило, до кількості навчальних прикладів, для яких виконується умова правила

$$\text{accuracy}(\alpha \to \beta) = \frac{\text{support}(\alpha \to \beta)}{\text{support}(\alpha)}.$$

3. Покриття, *coverage*($\alpha \to \beta$) – відношення кількості навчальних прикладів, для яких виконується все правило, до кількості навчальних прикладів, для яких виконується наслідок правила

$$\text{coverage}(\alpha \to \beta) = \frac{\text{support}(\alpha \to \beta)}{\text{support}(\beta)}.$$

4. Допоміжні показники, наприклад *coverage*(α) – частина навчальних прикладів, для яких виконується умова правила

$$\text{coverage}(\alpha) = \frac{\text{support}(\alpha)}{|U|}.$$

Результати класифікації нових прикладів подають *матрицею помилок*, у якій *True*(*X*) є кількістю прикладів, правильно віднесених класифікатором до класу *X*, а *False*(*X*) – неправильно. На основі цих оцінок розраховують коефіцієнт успішності (*КУ*) класифікації

$$КУ = \frac{\sum_{i=1}^{k} True(X_i)}{\sum_{i=1}^{k} True(X_i) + \sum_{i=1}^{k} False(X_i)},$$

де $k = |V_d|$ – кількість класів у таблиці прийняття рішень. Для покращення результату класифікації в процесі навчання додатково використовують методи *cross-validation*, *bootstrap* і *bagging* [9, 10].

Якість розрізнення класифікатором класів X_1 та X_2 можна оцінити з допомогою ROC-кривої (ROC, receiver operating characteristic) [11]. Ця крива показує поведінку класифікатора для різних порогових значень τ. У разі небінарної класифікації один з класів вважають класом X_1, а решту об'єднують у клас X_2. Для побудови ROC-кривої для кожного значення τ обчислюють матрицю помилок розмірів 2×2 та розраховують частину прикладів правильно (*TPR*, true positive rate) та неправильно (*FPR*, false positive rate) віднесених класифікатором до класу X_1 за формулами

$$TPR(\tau) = \frac{True(X_1)}{True(X_1) + False(X_2)}, \; FPR(\tau) = \frac{False(X_1)}{False(X_1) + True(X_2)}.$$

Частину правильно класифікованих прикладів відкладають на осі ординат, а неправильно – на осі абсцис. Площа *AUC* (area under curve) під ROC-кривою є показником якості розрізнення класифікатором прикладів класів X_1 та X_2. Значення $AUC = 0.5$ відповідає відсутності у класифікатора здатності розрізняти класи, а $AUC = 1$ означає ідеальну класифікаційну здатність [8].

Цілі дослідження. Мета досліджень полягала в аналізі таблиці даних з результатами психологічних тестів. Для цього з використанням наближених множин побудовано та порівняно декілька класифікаторів. Приділено особливу увагу питанням оцінювання якості класифікаторів. Побудовано та досліджено якість класифікаторів на основі наближених ($HF<1$) та точних редуктів ($HF=1$). Порівняння класифікаторів за допомогою ROC-кривих, значення *AUC* та коефіцієнта успішності виявило найкращий з них. Це дало змогу зробити висновки щодо якості оцінювання персоналу шляхом проведення тестувань і надати рекомендації для покращення наявної процедури прийняття рішень.

Експериментальна частина. Таблиця даних з результатами тестувань містить 188 прикладів та 38 атрибутів і не має невідомих значень. Домени умовних атрибутів містять дані про професію, місце праці, вік, досвід працівника, а також результати вимірювання швидкості реакції, об'єму пам'яті, здатності до концентрації, професійних навичок, мотивації тощо. Атрибутами прийняття рішень у таблиці є "профпридатність", яку розраховують психологи на основі результатів тестів, а також "успішність" та "надійність", які оцінюють керівники працівника, що проходить тестування. Оцінки v_1="відмінно", v_2="добре", v_3="задовільно" та v_4="незадовільно" є мітками класів X_i, до яких належать приклади таблиці, $X_i = \{x \in U \mid d(x) = vi\}$, $i = 1,...,4$. В процесі побудови класифікаторів навчальною множиною є 94 приклади, а решта 94 приклади утворюють тестову множину. Останню

використано для тестування та оцінювання якості класифікації. Для різних атрибутів прийняття рішення та різних значень *HF* проведено дві групи експериментів. У першій групі на основі 35 умовних атрибутів побудовано класифікатори №1–3 для атрибута прийняття рішення "профпридатність". У другій групі експериментів на основі 37 умовних атрибутів побудовано класифікатори №4–6 для атрибута прийняття рішення "надійність".

У табл. 1 наведено параметри класифікаторів і результати їх застосування для класифікації 94 тестових прикладів. З наведених даних видно, що для класифікатора №1 застосуванням алгоритму Джонсона кількість умовних атрибутів скорочена з 35 до 4, а для класифікатора №4 – з 37 до 5. У першій групі експериментів для заданого найменшого значення HF алгоритм зменшив кількість атрибутів редукта до двох – "вікова група" та "оцінка інтегрального показника". У другій групі експериментів до атрибутів редукта з найменшим значенням *HF* увійшли "місце роботи" та "оцінка обсягу уваги". Також видно, що зі зменшенням ступеня підтримки *HF* збільшується коефіцієнт успішності класифікації. Для класифікаторів №4–5 значення площі під ROC-кривою не наведено, оскільки через низьку успішність класифікації (0.05 та 0.35) ROC-крива не дозволяє отримати змістовні результати.

Таблиця 1

Якісні характеристики класифікаторів

Атрибут прийняття рішення	Номер класифікатора	Ступінь підтримки, HF	Кількість атрибутів редукта	Кількість правил	Коефіцієнт успішності, КУ	Площа, AUC
профпридатність	1	1.0	4	60	0.6	0.68
	2	0.99	3	32	0.78	-
	3	0.96	2	13	0.82	0.81
надійність	4	1.0	5	86	0.05	-
	5	0.98	3	45	0.35	-
	6	0.91	2	16	0.5	0.54

На основі інформації з табл. 1 побудовано графіки, зображені на рис. 1. Значення успішності класифікації для атрибута прийняття рішення "профпридатність" об’єднані однією лінією, а для атрибута "надійність" – іншою. Ці лінії показують, як залежить успішність класифікаторів від значення ступеня підтримки. З його збільшенням успішність класифікації зменшується від 0.82 до 0.68 у першій групі експериментів, і від 0.5 до 0.05 – у другій. Класифікатор №3 показує найкращу успішність класифікації – 0.82, що відповідає 82% правильно класифікованих тестових прикладів. У табл. 2 наведено результати роботи класифікатора №1. Цей класифікатор визначив

клас 72 тестових прикладів, з них 56 прикладів класифіковані правильно. Кількість правильно класифікованих прикладів подана на головній діагоналі таблиці для кожного з класів. Решта 22 приклади не класифіковано, оскільки класифікатор не містив правил, які виконуються для цих прикладів. У табл. 3 показано результати роботи класифікатора №3, який правильно класифікував 77 прикладів і не зміг класифікувати лише один приклад. Отже, 13 правил класифікатора №3, згенеровані на основі редукта з двома атрибутами, виконуються для більшої кількості прикладів і показують найкращий результат успішності класифікації. Ці правила є більш загальними, ніж 60 правил класифікатора №1, які побудовані на основі редукта з 4 атрибутами.

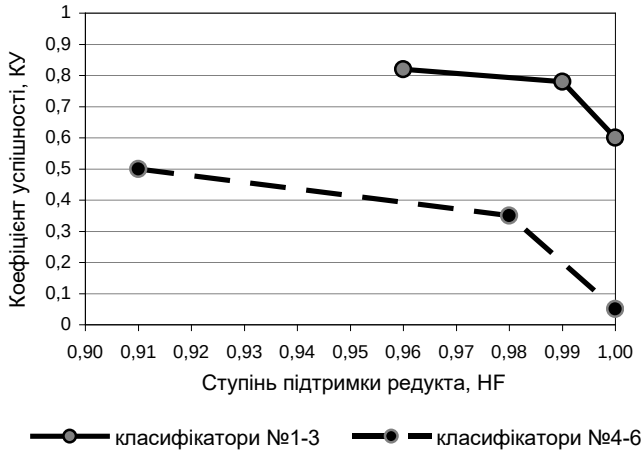

Рисунок 1 – Порівняння класифікаторів за показником КУ

Таблиця 2

Матриця помилок класифікатора №1

Справжній клас	Прогнозований клас			
	X_1	X_2	X_3	X_4
X_1	2	2	0	0
X_2	3	49	2	0
X_3	1	7	5	0
X_4	0	1	0	0

Класифікатори також порівняно за допомогою ROC-кривих, які показують співвідношення між кількостями прикладів, правильно і неправильно віднесених до класу X_2 за різних порогових значень τ. На рис. 2 зображено три ROC-криві для класифікаторів №1, №3 та №6, відповідно.

Таблиця 3

Матриця помилок класифікатора №3

Справжній клас	Прогнозований клас			
	X_1	X_2	X_3	X_4
X_1	1	4	0	0
X_2	3	66	1	0
X_3	0	6	10	1
X_4	0	0	1	0

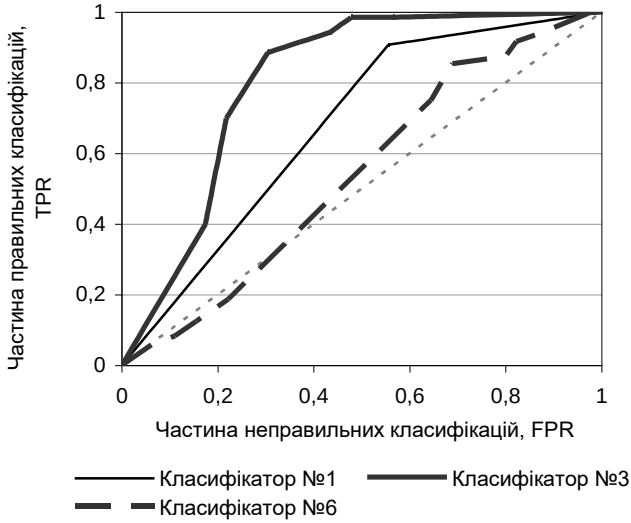

Рисунок 2 – ROC-криві, які показують здатність класифікаторів визначати клас X_2 з міткою "добре"

З рисунку видно, що класифікатор №3 краще від інших розпізнає приклади класу X_2: лінія класифікатора №3 домінує над лінією класифікатора №1 й має більше значення площі під кривою $AUC = 0.81$. Для лінії класифікатора №1 площа $AUC = 0.68$ (див. табл. 1). Найменшу класифікаційну здатність ($AUC = 0.54$) показує класифікатор №6, побудований для атрибуту прийняття рішення "надійність". Це можна пояснити тим, що значення цього атрибута визначають керівники працівника суб'єктивно і не використовують при цьому об'єктивних закономірностей, які містить таблиця з даними тестування. Результати, отримані для класифікаторів №4–6, підтверджують, що цей атрибут не впливає на залежності у таблиці даних. Ці класифікатори, побудовані для атрибута прийняття рішення "надійність", у найкращому випадку показують

успішність класифікації лише 50%, що не можна вважати прийнятним результатом. Найкращі результати показали класифікатори №1–3, які побудовано для атрибута прийняття рішення "профпридатність". Це свідчить про те, що цей показник отримано на основі якісно підготовлених даних, які містили реальні закономірності. Такі закономірності були виявлені під час аналізу даних, вони дозволили побудувати класифікатори, результати застосування яких до тестових прикладів показали хороше узгодження із рішеннями, що були прийняті психологами.

Висновки. Результати проведених досліджень показали, що використання наближених множин у задачах класифікації є виправданим для побудови класифікаторів та прийняття рішень на їх основі. Частина опрацьованих даних була отримана в результаті суб'єктивного оцінювання ситуацій. З допомогою якісних характеристик класифікаторів вдалось оцінити вплив суб'єктивного фактора в даних на прийняття рішень. Це надає можливості для реорганізації методики проведення тестування, у якій потрібно зменшити суб'єктивний вплив на результати прийняття рішень. Показано, що результати навчання можна покращити зменшенням кількості атрибутів і побудовою наближених редуктів, на основі яких формулюють правила прийняття рішень. В проведених експериментах досягнуто 82% успішності класифікації тестових прикладів зі зменшенням кількості атрибутів з 37 до 2. Порівняння класифікаторів за допомогою ROC-кривих та за значеннями параметрів *AUC* і *КУ* дозволило оцінити їхню якість на тестових прикладах та виявити серед них класифікатор №3 як найкращий. Проведені дослідження також дозволили виявити серед атрибутів прийняття рішення атрибут "профпридатність", значення якого найбільше враховує залежності в даних. Отримані результати використано для покращення методики тестування та моніторингу роботи персоналу підприємств енергетичної галузі. Для забезпечення більшої незалежності результатів від вибору навчальних і тестових прикладів надалі варто застосувати додаткові методи навчання і тестування класифікаторів.

Список літератури: ***1.*** Mitra S., Pal S. K., Mitra P. Data mining in soft computing framework: a survey // IEEE Transactions on Neural Networks. − 2002. − Vol. 13. – P. 3–14. ***2.*** Wang G., Liu Q., Yao Y., Skowron A. (Eds.). Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. // Proc. of 9th Int. Conference, RSFDGrC-2003. − Springer. − 2003. ***3.*** Заде Л. А. Роль мягких вычислений и нечеткой логики в понимании, конструировании и развитии информационных/интеллектуальных систем. // Новости Искусственного Интеллекта. РАИИ. – 2001. – №2–3. С. 7–11. ***4.*** Komorowski J., Polkowski L., Skowron A. Rough Sets: A Tutorial. // Eds. S. K. Pal and A. Skowron. Rough Fuzzy Hybridization: A New Trend in Decision-Making. − Springer-Verlag. − 1998. – P. 3–98. ***5.*** Polkowski L. Rough Sets: Mathematical Foundations. – Physica-Verlag. − Heidelberg. − 2002. ***6.*** Pawlak Z. Rough Sets – Theoretical Aspects of Reasoning about Data. − Kluwer Academic Publishers. − Dordrecht. − 1991. ***7.*** Øhrn A. ROSETTA Technical Reference Manual. − 2001. http://www.idi.ntnu.no/~aleks/. ***8.*** Øhrn A. Discernibility and Rough Sets in Medicine: Tools and Applications, PhD thesis. Norwegian Univ. of Science and Technology, Dep. of Computer and Information Science. − 2000. ***9.*** Efron B., Tibshirani R. An Introduction to the Bootstrap. − Chapman Hall. − New York. − 1993. ***10.*** Bauer E., Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants.

// Machine Learning. − Springer Netherlands. − 1999. − №36. – P. 105–139. ***11.*** Hastie T., Tibshirani R., Friedman J. The Elements of Statistic Learning; Data Mining, Inference, and Predicting. – Springer-Verlag. –2001.

Поступила в редколлегию 26.11.07

УДК 007.681

О. Я. ЛАЗАРЕВА, канд. техн. наук

МЕТОДИКА АВТОМАТИЗАЦИИ ФОРМИРОВАНИЯ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЕЙ

В статті пропонується один з підходів до створення термінологічних словників. Основою реєстру термінологічного словника може бути частотний список словосполучень певної граматичної структури, автоматично виділених з текстів деякої предметної галузі. Напевно, що такий словник буде надмірним і буде потребувати коригування експертом.

В статье предлагается один из подходов к созданию терминологических словарей. Основой словника терминологического словаря может служить частотный список словосочетаний определенной грамматической структуры, автоматически выделенных из текстов определенной предметной области. Естественно, что такой словарь будет избыточным и потребует корректирования экспертом.

The paper suggests an approach to creation of the vocabulary of terminological dictionaries. Word combinations of definite grammar structures being automatically extracted from the texts of some domain and arranged in a frequency list may constitute the basis for a terminological dictionary. The obtained vocabulary is sure to be excessive and should be edited by an expert.

Современное состояние «инфосферы», пронизывающей все виды человеческой деятельности, характеризуется лавинообразным увеличением потоков информации, все большим слиянием и взаимопроникновением отдельных областей знаний, что вызывает потребность в создании и обновлении словарей, обслуживающих данные предметные области. В связи с этим возникает необходимость пересмотра подходов к самому процессу создания или пополнения словарей. Традиционно специалисты-предметники, часто в сотрудничестве с лингвистами, составляли такие словари - и словари на бумажных носителях, и словари, использующиеся в автоматизированных информационных системах (классификаторы, рубрикаторы, информационно-поисковые тезаурусы) - на основе анализа достаточно большой коллекции документов по некоторой тематике. И совершенно очевидно, что сам процесс накопления массива лексических единиц был и остается чрезвычайно трудоемкой операцией. Кроме того, учитывая стремительное развитие практически всех отраслей науки и технологий, с одной стороны, и достаточно длительный процесс подготовки или модернизации словарей, с